Which One To Read? Factors Influencing the Usefulness of Online Reviews for RE

Ben Charrada, Eya

DOI: https://doi.org/10.1109/REW.2016.022

Which One To Read? Factors Influencing the Usefulness of Online Reviews for RE

Eya Ben Charrada
Department of Informatics
University of Zurich
Switzerland
Email: charrada@ifi.uzh.ch

***Abstract*—Reviews for software products contain much information about the users' requirements and preferences, which can be very useful to the requirements engineer. However, taking advantage of this information is not easy due to the large and overwhelming number of reviews that is posted in various channels. Machine learning and opinion mining techniques have therefore been used to process the reviews automatically and to generate summaries of the data to the requirements engineer. However, one of the important challenges for these techniques lies in how to automatically assess the relevance of the reviews for the requirements engineer. So far, most techniques use intuition-based criteria for this task. In this work, we collect and present a list of factors that were found to impact the helpfulness of product reviews for customers. We then discuss to what extent these factors are likely to impact the usefulness of reviews for requirements engineering tasks. The factors can be used to support the automated identification of relevant reviews.**

I. Introduction

A large number of online reviews are available for software products over various platforms and channels such as social channels, stores for mobile apps, forum discussions, etc. These reviews include feedback that can be used to support several requirements and software engineering tasks [1][2]. However, the challenge with using online reviews is that they are too numerous and overwhelming for the human mind and it is therefore difficult to use them efficiently. In fact, "a wealth of information creates a poverty of attention and a need to allocate that attention efficiently among the overabundance of information sources".[1] In an attempt to address this challenge, researchers proposed to use machine learning and opinion mining approaches to automatically analyse large numbers of reviews for products (e.g. [3][4]). Recently, these techniques have been attracting much attention from the software engineering and the requirements engineering communities [1][5]. For instance, several approaches have been developed to classify and summarize customer feedback obtained from various channels (e.g. [6][7]). The output quality of the review processing techniques depends on their capability to identify the reviews that are most useful among the abundance of information. So far, most developed approaches use criteria that are based on intuition.

In this work, we collect, via a literature survey, and present a list of factors that were found to influence the helpfulness of reviews for customers. We also discuss to what extent these factors are likely to apply for requirements engineering tasks. The output of this work can therefore be used to support the automated identification of relevant reviews for requirements engineers.

The remainder of the paper is structured as follows. In the next section, we present how machine learning is being used to analyse customer reviews and feedback. In Section III, we explain why customers and software engineers are in many cases interested in similar information from the reviews. We present the factors that influence the helpfulness of reviews for customers in Section IV and discuss their relevance to requirements engineering in Section V. The limitations and threats are discussed in Section IV. Finally, we present future work directions and conclude in Section VII.

II. Machine Learning for Analysing and Prioritizing Reviews

Several approaches have been developed to process data coming from various channels such as app stores (e.g. [6]), online sellers and websites offering review possibilities (e.g. [8]), blogs (e.g. [9]) or Twitter [10][11]. Most of the approaches focus on classifying reviews into categories (e.g. [12][13]) according to their type (simple praise, feature request, reporting bug, etc.) and counting frequencies and occurrences of features and sentiments to give the software engineer a high-level overview of the main topics addressed in the reviews and the users' opinions about them (e.g. [6][14][7]). Approaches for identifying the most relevant reviews and/or prioritizing reviews among each other are still scarce. Chen at al. [15] is among the few works that gave prioritization a special focus and did a first interesting attempt to identify relevant reviews using criteria they defined from observing some developer forums combined with criteria based on intuition. The works of Villarroel et al. [16] and Keertipati et al. [17] also tried addressing the prioritization problem. For this, Villarroel et al. [16] used criteria based on intuition [16] while Keertipati et al. [17] used a combination of criteria already in use by researchers (frequency and rating) and criteria they had derived from psychology literature (emotions and deontics). Identifying relevant reviews, however, is still a challenge that requires more focus and effort from requirements engineering researchers [5]. To the best of our knowledge there is still no

[1]A quote from Herbert A. Simon

work that focuses on studying the characteristics of relevant and useful reviews for requirements engineering.

III. Different but Similar: the Customer vs the Requirements Engineer

Customers and requirements engineers look at reviews for different purposes. Customers usually look for information that will support them deciding about whether or not to get the product or service. Requirements engineers, however, look for information that will support them improving and evolving the software. Although the purposes are different, it is likely that they are interested in the same type of information. For example, information about the advantages and disadvantages of the product is relevant for both purposes. When considering the topics that reviews cover, as reported by Pagano and Maalej [18], it is clear that most of the frequent topics (praise, helpfulness, feature information, shortcoming, comparing the application to other ones) are of relevance for both the customer and the requirements engineer. Due to this overlap in the topics of interest, we conjecture that the factors influencing the helpfulness of reviews for customers also apply, to a certain extent, for the requirements engineer. The only study we found on review usefulness from the designers' perspective is the work of liu et al. [19], which is based on an experiment with undergraduate students in product engineering. The properties that the student participants reported about useful reviews are: long review covering the reviewer's preferences, mentioning many different features, pointing out the likes and dislikes of the product, and comparing the product to another product. Since such information is similar to what customers are interested in, this provides additional support to our conjecture.

IV. Factors Influencing Review Helpfulness in Literature

Several studies have been conducted to identify the factors influencing review helpfulness for customers. Our goal is to collect these factors by surveying the literature and discuss to what extent these are likely to apply in a requirements engineering context. In this section, we present how we surveyed the studies and report the factors we found. The discussion of the applicability of these factors to requirements engineering is presented in Section V.

A. Paper Selection

The selection of the papers was done as follows. We did a search on scopus with the following keywords:

("online reviews" OR "product reviews" OR "user reviews" OR "customer reviews" OR "consumer reviews")
AND
("useful" OR "helpful" OR "usefulness" OR "helpfulness")

We performed a first filtering of the papers based on the number of citations as displayed by Scopus. For all papers that were published prior to 2016, we only included those that had a number of citations that is above a certain threshold. The older the publication date the higher the threshold we set (3 citations for 2015, 7 citations for 2014, 15 citations for 2013, etc.). For 2016, we considered all the papers that were published when we did the search (early June) without restrictions on the number of citations. We then performed a second filtering based on the topic and focus of the papers. All papers that did not relate to the topic of user reviews for goods were ignored. We also filtered publications that propose approaches for automated review analysis without studying the factors influencing the helpfulness of the reviews. This led to 24 papers that we identified as relevant to the topic. One of the papers ([19]) had a slightly different topic and has already been mentioned in the previous section. For one article, we could not get access to the full text and, thus, did not include it. We used the other 22 papers to identify the factors that were found to have an influence on review helpfulness.

TABLE I
Summary of the factors influencing the helpfulness of reviews for customers. The influence can be positive, negative or mixed.

Category	Factors
Language	Readability [20][21][22] Elaborated sentences [23] Sophisticated words [23] Proportion of negative words [24] Number of one-letter words [25] Negative style characteristics [26] Spelling errors [21]
Volume and longevity	Length [27][28][26][29][25][30][31][23][24][32] Average number of words in sentence [25] Image count [31] Longevity [28]
Rating, sentiment and emotions	Negative [33][23][34][35] Positive [34][22][26][35][32][29] Rating extremity [33][30] Parallel with the majority average rating [24] Neutral polarity in text & Sentiment [28] Emotions [36]
Content	Concrete [37] Argumentation [34] Mixture of objective and subjective elements [21] Explained actions [38] Explained reactions [38] Proportion of product-descriptive statements [26] Proportion of reviewer-descriptive statements [26]
Reviewer	Reputation and rank [27][24][23] Helpfulness of previous reviews [21][29] Identity information disclosure [27][39] Number of followers [31] Positive historical record [20] Level of expertise [31] Customer vs expert [37][34] Expressed innovativeness [32]

B. Factors

We identified 32 factors which we classified into five categories that are presented in Table I.

1) Language: Several factors relating to language properties were found to influence the helpfulness of reviews. Fang et al. [20] found that a precise and easy to understand writing style had a positive impact on the perceived

helpfulness. This result is consistent with previous studies, which found that increased text readability has a positive impact on helpfulness [21] and that the readability effect is greater than the effect of length [22]. Lee and Choeh [25] found that the number of one-letter words also relates to increased helpfulness [25] and attributed this to higher readability. An increase in spelling errors was found to negatively impact review helpfulness [21] and a greater use of negative style characteristics (e.g. misspellings, bad grammar, slang-inexpressiveness, repetition) was associated with lower-value reviews [26]. The use of elaborated sentences and sophisticated words seems to interest consumers and lead them to read and vote more for these reviews than for the ones written in too simple text ("e.g., those composed of short or even broken sentences using simple vocabularies") since they may appear unprofessional [23]. Furthermore, the frequent use of negative words in the review message was connected to an increase in review helpfulness [24]. This could be attributed to the relation between negative reviews and helpfulness, which we present later.

2) Volume and longevity: Length (also called depth) is the characteristic that has been studied the most within the papers we identified. Most studies consistently found that the length of reviews (word count) positively impacts helpfulness [27][28][25][30][29][31][24][32]. This relation holds until a certain length threshold (1000 to 1500 words according to [24]) which seems to depend on the type of product. After the threshold, the helpfulness starts decreasing [24][23][29], which could be due to the readers disregarding very long reviews [23]. The effect of length on helpfulness has been identified for various types of products and seems to have greater effect for search goods[2] than for experience goods[3] [30][32]. Some studies considered more developed notions of length. For example, in order to ignore repetition, Qazi et al. [40] considered the number of concepts in a review and the average number of concepts in a sentence. They found that both factors had a positive impact on helpfulness. Still, reviews that are too long are not useful. Schindler and Bickart. [26] considered the number of statements in a review, which they also found to be associated to high-value reviews but only up to a certain point. The average number of words in sentences was also found to positively impact helpfulness [25], which the authors attribute to higher informativeness. Also related to the quantity of information, a study based on data from restaurant review websites found that the image count positively impacts helpfulness [31].

Salehan and Kim. [28] found a positive relation between longevity and helpfulness which means that older reviews are perceived as more helpful than newer ones. However, since the study was based on the votes received for reviews on Amazon, this might be explained by the way Amazon sorts the reviews, which is based on votes. This means that newer reviews are likely to stay at the end of the list and are likely to receive less attention [28].

3) Rating, sentiment and emotions: Several studies have investigated the relation between the rating given by the reviewer and the helpfulness of the review. The output seems to differ much from one study to another. Negative reviews and ratings were found to be perceived as more helpful in [33][23]. This hypothesis was however not supported by the study of Baek et al. [24] and was even contradicted by the study of Pan and Zhang [32] who found that positive reviews have a greater probability of being rated helpful than negative ones and that the helpfulness decreases with decreased rating. This is also supported by the study of Korfiatis et al. [22] who found that helpfulness is affected by the positive rating value, by the study of Huang et al. [29] who found that for top reviewers, positive reviews are more likely to be helpful, and by the study of Schindler and Bickart. [26] who found that greater proportions of positive statements are associated with high value reviews.

Willemsen et al. [34] found that the valence was contingent on the type of product: negative reviews were more useful for experience products while positive reviews were more useful for search products. However, an opposite relation was suggested by Sen and Lerman [35] who found that there is a negativity bias for utilitarian products.[4] This is because the readers are likely to attribute the reviewer's negative opinion to product-related reasons. For hedonic products,[5] however, the readers attribute the negative opinion to the reviewer's internal (non product-related) and is therefore less likely to be useful. In an attempt to study the relation from a different perspective, Lee and Koo [41] found that review valence, alone, does not affect review usefulness but that it depends on the consumers' regulatory mode orientation: consumers high in assessment[6] prefer negative reviews while customers high in locomotion[7] prefer positive reviews.

Regarding the extremity of the rating, Mudambi and Schuff. [30] found that reviews with extreme ratings are less helpful than those with moderate ones for experience goods. Similarly, Kuan et al. [23] found that review extremity has a negative effect on helpfulness. However the findings of Park and Nicolau [33] are different since they found that extreme ratings are perceived as more helpful and enjoyable than moderate ratings in the context of restaurant reviews.

When studying the influence of polarity on helpfulness, Salehan and Kim [28] found that reviews that include sentiments and have a neutral polarity (i.e. similar number of positive and negative sentiments are in the review) are more

[2]Search goods are "those for which consumers have the ability to obtain information on product quality prior to purchase" [30] e.g. cameras.

[3]Experience goods are "products that require sampling or purchase in order to evaluate product quality" [30], e.g. music.

[4]"Utilitarian products are primarily cognitive and functional and are purchased out of necessity, as a means to an end (e.g., batteries)" [38]

[5]"Hedonic products are primarily emotional and sensory and purchased out of desire, for the intrinsic enjoyment they provide (e.g., video games)" [38]

[6]In regulatory mode theory, assessment is about "critically evaluating entities or states (such as goals or means) against alternatives in order to judge their relative quality" [41]

[7]Locomotion is about "moving among states and committing the psychological resources needed to initiate and maintain goal-related movement straightforwardly, without distractions or delays" [41]

helpful. This could be explained by the fact that reviews that are only positive or only negative are considered as biased by consumers and thus are perceived as less helpful [28]. Also related to rating, Baek et al. [24] found that the reviews are most helpful when they are in parallel with the majority average rating and that the higher the divergence from product average rating, the lower the review helpfulness.

Yin et al. [36] studied the role of two particular types of emotions: anxiety and anger. They found that reviews containing content indicative of anxiety were perceived as more helpful than those expressing anger. They also found support to the hypothesis that this is mediated by the perceived cognitive effort. A possible explanation that the authors give is based on the effect of anxiety and anger on individuals: anxious individuals are likely to devote more cognitive effort to the reviewing task, while angry individuals are likely to engage in processing that requires less thought. Furthermore, these emotions can also influence how the readers perceive the level of cognitive effort invested by the writer of the anxious and angry reviews.

4) Content: Reviews that are concrete are found to be more helpful than abstract ones [37]. The argumentation included to support the opinion presented in the review, both in terms of density and diversity, was also found to be a significant predictor of review helpfulness [34]. Furthermore, reviews that include only objective or only subjective elements are perceived as less helpful than those including both elements [21]. In terms of quantity of objective and subjective elements, customers were found to prefer more subjective elements for experience goods and more objective elements for search goods. Moreover, Moore [38] found that for utilitarian products, reviews that include explained actions (e.g. "I chose this product because...") are more helpful for customers, while for hedonic products reviews are more helpful when they include explained reactions (e.g. "I love this product because..."). Schindler and Bickart [26] found that a greater proportion of product descriptive statements and a greater proportion of reviewer-descriptive statements (but only up to point) were also associated with high-value reviews.

Pan and Zhang [32] found that the reviewer's expressed innovativeness, which they define as "the predisposition towards new products as revealed in the review content" has an inverted U effect on the perceived usefulness. This means that reviews with medium innovativeness are more helpful than the most and least innovative ones.

5) Reviewer: Several factors relating to the reviewer have been studied in the context of review helpfulness. For instance, the reviewer's reputation [27][24][23], helpfulness of previous reviews [21][29] as well as the number of followers [31] were found to have a positive influence on the review usefulness. Regarding the disclosure of the identity of the reviewer, inconsistent results were found. In fact, while some researchers found that the disclosure of identity-descriptive information increases helpfulness and trustworthiness [27][39] other researchers found that mere exposure of the reviewer's real name was not enough to increase helpfulness [24].

TABLE II
FACTOR RELEVANCE TO REQUIREMENTS ENGINEERING

	Language	Volume	Rating	Content	Reviewer (1 & 2)*
Assessment	Low	Low	High	Low	1:Low, 2:High
Elaboration	High	High	Low	High	1:High, 2:High

*We consider information about the reviewer as a reviewer (1) and as a user (2).

The difference between the reviews written by customers and experts were also studied. Cheng and Ho [31] found that the level of expertise had a positive impact on usefulness, while Willemsen et al. [34] found that when the expertise level is based on self claims, the influence on helpfulness was weak. Li et al. [37] also found that customers consider customer-written reviews to be more useful than the reviews written by experts although these are usually in-depth and unbiased. In the context of tourism, Fang et al. [20] found that reviews that are written by reviewers who have a positive historical record of reviewers are likely to be more useful.

V. APPLICABILITY TO REQUIREMENTS ENGINEERING

A. Online Reviews for RE

To study the relevance of the factors in the context of requirements engineering, we consider two tasks that a requirements engineer could perform with the support of reviews: (1) assessment and (2) elaboration.[8] In an assessment task, the requirements engineer tries to get a high-level overview of the stakeholders' requirements, problems and preferences in order to make informed decisions about what requirements or misbehavior to focus on next and possibly prepare a light-weight description of those. For this task, quantitative data representing what stakeholders like, dislike and request in the software is very relevant. After selecting the requirements of interest, the requirements engineer starts the elaboration task during which the requirements are explored more in-depth and are specified with enough details to be implemented. For this task, relevant qualitative data is essential. In the rest of this section, we discuss the relevance of the identified factors for each of these tasks. A summary of the factors' relevance is presented in table II.

B. Relevance of the factors

1) Language: Language properties are likely to play an important role with regard to the usefulness of reviews for elaboration tasks. For instance, when trying to grasp the details of a requirement from reviews, it is extremely important for these reviews to be clear and easy to read. In this context, non-understandable reviews might be more harmful than beneficial since they are likely to waste much cognitive effort and eventually cause frustration. Furthermore, when the review has good language properties (e.g. readable, less spelling errors, elaborated sentences) this could indicate that the reviewer invested time and cognitive effort to write a good review while

[8]These tasks are inspired from and similar to the steps for just-in-time requirements analysis [42] where requirements are first sketched and then elaborated during development.

keeping the reader in mind. Consequently, it is not surprising to find content that is well thought and more useful from this type of reviews than from reviews with bad readability. For assessment tasks, however, language properties are much less important since the requirements engineer is more interested in quantitative and high-level data and would, hence, not look at the detailed text of the reviews.

2) Volume and longevity: When considering the customer's perspective, data volume was consistently found to have a positive effect on helpfulness at least until a certain point. We expect data volume to also increase usefulness for the requirements engineer during the elaboration phase, for three reasons. First, similarly to the language properties, length could reflect the time and effort that the reviewer invested in writing, and would, hence, positively influence their quality and usefulness. Second, length can play a key role in reducing misunderstandings in the context of eliciting requirements from reviews. In fact, since the requirements engineer usually has no previous communication or contact with the reviewers, the *implicit shared understanding* [43] between the engineer and the reviewer is very limited. Consequently, there is a need for a minimum length to establish a certain degree of shared understanding among them and reduce misunderstanding. Third, when elaborating and refining a requirement, details are of high importance. Obviously, details are more likely to be present in long reviews than in short ones.

Although the helpfulness of reviews as perceived by customers seems to decrease after a certain length threshold, this is not likely to be the case for requirements engineers. In fact, customers usually aim at reading many reviews to get an overall overview of the product and also to check the consistency of the reviews among each other. Therefore, they might disregard reviews that are too long since they require time and cognitive effort that is beyond what they would like to invest in a single review. This is, however, different for a requirements engineer who is performing an elaboration task. Indeed, for such a task, identifying the few reviews that provide detailed information about the requirement of interest will be of a greater help to the engineer than looking at numerous reviews that are reporting about the requirement on a high-level. The threshold, however, might apply during the assessment phase since the requirements engineer needs to look at various concrete reviews in order to informally sketch the requirement. In the case where the requirements engineer is interested in purely quantitative data, then review length is not likely to play any role on relevance.

Regarding the data format, screenshots and images could be of high relevance for elaboration tasks since they can help the engineer to grasp the idea behind the request or to understand the details of a reported misbehaviour.

With respect to longevity, it might be interesting for elaboration tasks to consider older reviews that have been updated after posting since this gives information about how and why a user's opinion evolved over time. Nevertheless, we do not think that longevity alone is a positive indicator about review usefulness for software products since these are likely to evolve very quickly and thus reviews can rapidly become obsolete if not updated regularly. Old reviews can, nonetheless, be used to study how the users' preferences evolve over time, similarly to what Chen et al. [15] proposed.

3) Rating, sentiment and emotions: Results about the influence of rating and sentiment on review helpfulness for customers are very inconsistent. For the requirements engineer, rating and sentiment, alone, are not enough to predict the usefulness of reviews. However, depending on what the requirements engineer is looking for, rating and sentiment can be used for narrowing the search. Specifically, if the requirements engineer is interested in what features the users like most, then such information is more likely to be in positive reviews. When looking for information about software misbehaviours, then directing the search towards negative reviews could be more helpful. Rating and sentiment are likely to be more relevant for assessment tasks than for elaboration tasks since for elaboration tasks the engineer would be interested in any review that gives details about the requirements of interest regardless of whether the review is positive or not.

So far, researchers have given more weight to negative reviews than to positive ones when performing prioritization (e.g. [15][16]). Although this reflects the natural human behaviour, which prefers avoiding losses to acquiring gains, the relation between the rating/sentiment and the review relevance is likely to be much more complex especially when looking for qualitative data. In fact, it is plausible that reviews that are too negative are not constructive and are, thus, less useful than positive ones for the elaboration task. Similarly, reviews that are too positive could reflect user bias towards a product and would also not be so useful. Consequently, we expect reviews that have moderate rating and neutral polarity to be the least biased and thus be the most useful for the requirements engineer during the elaboration task.

With respect to emotions, they are likely to be of interest when looking for qualitative data (elaboration) since they can reflect the cognitive effort that the reviewer invested in writing the review.

4) Content: Most factors in the content category are likely to be of relevance for the elaboration task. In particular, detailed and concrete reviews that include argumentation of the reviewer's point of view are more likely to be useful than abstract ones. Furthermore, several content elements (namely concreteness, argumentation, explanations of the reviewer's actions and/or reactions and inclusion of statements describing the product and the reviewer) are also essential to ensure a minimum shared understanding and thus reduce the risk of the requirements engineer misunderstanding the reviewer's point of view. Statements describing the reviewer can also be used to identify groups of users and stakeholders, which is of high importance for the assessment and the elaboration tasks.

5) Reviewer: There are two types of information about reviewers that can be highly relevant in a requirements engineering context. The first type of information is about the reviewer's capabilities and skills for review writing. For example, when the reviewer has a record of reviews that were

perceived as helpful, then this can indicate that the reviewer is experienced in writing good reviews and we would therefore expect his/her new reviews to be helpful when looking for qualitative information. Reviewers' expertise is not likely to be important when looking for quantitative data though. The second type of information is about the reviewer as a user. This includes information about the reviewer's profile. This information is important for identifying groups of stakeholders and studying the behaviour, preferences and requirements for each of these groups separately. This also allows focusing on the requirements of the groups that are most interesting for the product owner. Stakeholder classification is relevant for both the assessment and the elaboration tasks.

Reviewer's record can also be useful for other purposes. For example, the historical record of reviewers can be consulted by the requirements engineer to know more about their background, preference and writing style and thus limit the risks of misunderstanding for elaboration tasks. Furthermore, the record could give information about the reviewers' personality. For instance, a positive record indicates that the reviewer is not constantly unsatisfied, but is more likely to be constructive.

C. Discussion

Approaches for automatically processing reviews need criteria to assess review relevance for the tasks that the requirements engineer is performing. Many factors presented in this paper can be used as criteria for relevance assessment. A good assessment of review relevance allows (1) filtering out irrelevant ones for generating quantitative results, (2) detecting the most relevant reviews for qualitative analysis and (3) sorting the reviews based on their relevance to a certain task. The criteria to be used are highly dependent on the task that the requirements engineer is performing. For example, when the goal is to elaborate a requirement to be implemented, length and readability of reviews are likely to be better predictors for usefulness than rating. Furthermore, a single criterion is not enough to assess the relevance of a review. Therefore, there is a need to combine several criteria in order to obtain a good assessment. Although many of the criteria can be easily checked automatically (e.g. length), some criteria are difficult to assess (e.g. concreteness) and could therefore be of limited use for automated analysis.

VI. Limitation and Threats

The usefulness of a review is likely to depend on many inter-related factors that are difficult to isolate and study. This explains the inconsistencies in some results among the studies and is a limitation of them. Customer bias also represents a threat for such studies. In fact, there are many factors which can influence how the customer perceives the helpfulness of a review. For example, Yin et al. [44] reported a consumer *confirmation bias* where consumers construct an initial belief/assessment about/of the product and then rate the reviews that confirm (or contradict) that belief as more (or less) helpful. This bias can lead to perceive positive (or negative) reviews as more helpful for products with high (or low) average ratings [44]. Such a bias could be aggravated by a reporting bias. In fact, reviews are mostly positive on average [30] and the ratings are influenced by the price of the product [45]. This results in the customers having a more or less positive initial belief which would then increase their perceived helpfulness for reviews with the same polarity. Furthermore, factors that give a positive or negative first impression (e.g. readability or spelling error) could also bias how the customer perceives the usefulness of the reviews. Fraudulent reviews pose another threat to the studies. For instance, positive review manipulation has been reported to be used for promoting lower quality products [46] while negative review fraud was found to increase with competition [47]. The reported studies did not consider the impact of fraudulent reviews on their results since it is extremely difficult or impossible to identify and filter out this type of reviews.

VII. Conclusion and Future Work

In this work, we surveyed and reported a list of 32 factors that impact the helpfulness of reviews from the customer's perspective. We grouped these factors into five categories that relate to (1) the language properties, (2) the data volume and longevity, (3) the rating, sentiment and emotions (4) the content properties, and (5) the reviewer. We also discussed to what extent are these factors likely to impact the usefulness of reviews when considering the perspective of the requirements engineer. This work is a first step towards supporting the automated assessment of review relevance for requirements engineering tasks. For future work, we first plan to validate and complement the list of factors presented here by involving requirements engineering practitioners and exploring what makes a useful review for them. This will be done via a combination of qualitative and quantitative approaches: interviews and practitioner survey. Then, we plan to conduct experiments to thoroughly assess the importance of some selected factors. In these experiments, practitioners will rate the usefulness of reviews for certain tasks, and we will explore to what extent the selected factors impact the perceived usefulness. Another direction for future work is to explore how the factors can be used to automatically order reviews based on their usefulness for the requirements engineer.

References

[1] W. Maalej, M. Nayebi, T. Johann, and G. Ruhe, "Toward data-driven requirements engineering," *IEEE Software*, vol. 33, no. 1, pp. 48–54, 2016.

[2] C. Iacob and R. Harrison, "Retrieving and analyzing mobile apps feature requests from online reviews," in *Mining Software Repositories (MSR), 2013 10th IEEE Working Conference on*. IEEE, 2013, pp. 41–44.

[3] A.-M. Popescu and O. Etzioni, "Extracting product features and opinions from reviews," in *Natural language processing and text mining*. Springer, 2007, pp. 9–28.

[4] M. Hu and B. Liu, "Mining and summarizing customer reviews," in *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 2004, pp. 168–177.

[5] M. Nagappan and E. Shihab, "Future trends in software engineering research for mobile apps," in *2016 IEEE 23rd International Conference on Software Analysis, Evolution, and Reengineering (SANER)*, vol. 5. IEEE, 2016, pp. 21–32.

[6] E. Guzman and W. Maalej, "How do users like this feature? a fine grained sentiment analysis of app reviews," in *Requirements Engineering Conference (RE), 2014 IEEE 22nd International*. IEEE, 2014, pp. 153–162.

[7] P. M. Vu, T. T. Nguyen, H. V. Pham, and T. T. Nguyen, "Mining user opinions in mobile app reviews: A keyword-based approach (t)," in *Automated Software Engineering (ASE), 2015 30th IEEE/ACM International Conference on*. IEEE, 2015, pp. 749–759.

[8] G. Ganu, Y. Kakodkar, and A. Marian, "Improving the quality of predictions using textual information in online user reviews," *Information Systems*, vol. 38, no. 1, pp. 1–15, 2013.

[9] M. Chau and J. Xu, "Business intelligence in blogs: Understanding consumer interactions and communities," *MIS quarterly*, vol. 36, no. 4, pp. 1189–1216, 2012.

[10] E. Guzman, R. Alkadhi, and N. Seyff, "A needle in a haystack: What do twitter users say about software? (to appear)," in *Requirements Engineering Conference (RE), 2016 IEEE 24nd International*. IEEE, 2016.

[11] A. Sharma, Y. Tian, and D. Lo, "What's hot in software engineering twitter space?" in *Software Maintenance and Evolution (ICSME), 2015 IEEE International Conference on*. IEEE, 2015, pp. 541–545.

[12] S. Panichella, A. Di Sorbo, E. Guzman, C. A. Visaggio, G. Canfora, and H. C. Gall, "How can i improve my app? classifying user reviews for software maintenance and evolution," in *Software Maintenance and Evolution (ICSME), 2015 IEEE International Conference on*. IEEE, 2015, pp. 281–290.

[13] W. Maalej and H. Nabil, "Bug report, feature request, or simply praise? on automatically classifying app reviews," in *Requirements Engineering Conference (RE), 2015 IEEE 23rd International*. IEEE, 2015, pp. 116–125.

[14] X. Gu and S. Kim, "" what parts of your apps are loved by users?"," in *Automated Software Engineering (ASE), 2015 30th IEEE/ACM International Conference on*. IEEE, 2015, pp. 760–770.

[15] N. Chen, J. Lin, S. C. Hoi, X. Xiao, and B. Zhang, "AR-miner: mining informative reviews for developers from mobile app marketplace," in *Proceedings of the 36th International Conference on Software Engineering*. ACM, 2014, pp. 767–778.

[16] L. Villarroel, G. Bavota, B. Russo, R. Oliveto, and M. Di Penta, "Release planning of mobile apps based on user reviews," in *Proceedings of the 38th International Conference on Software Engineering*. ACM, 2016, pp. 14–24.

[17] S. Keertipati, B. T. R. Savarimuthu, and S. A. Licorish, "Approaches for prioritizing feature improvements extracted from app reviews," in *Proceedings of the 20th International Conference on Evaluation and Assessment in Software Engineering*. ACM, 2016, p. 33.

[18] D. Pagano and W. Maalej, "User feedback in the appstore: An empirical study," in *Requirements Engineering Conference (RE), 2013 21st IEEE International*. IEEE, 2013, pp. 125–134.

[19] Y. Liu, J. Jin, P. Ji, J. A. Harding, and R. Y. Fung, "Identifying helpful online reviews: a product designer's perspective," *Computer-Aided Design*, vol. 45, no. 2, pp. 180–194, 2013.

[20] B. Fang, Q. Ye, D. Kucukusta, and R. Law, "Analysis of the perceived value of online tourism reviews: influence of readability and reviewer characteristics," *Tourism Management*, vol. 52, pp. 498–506, 2016.

[21] A. Ghose and P. G. Ipeirotis, "Estimating the helpfulness and economic impact of product reviews: Mining text and reviewer characteristics," *Knowledge and Data Engineering, IEEE Transactions on*, vol. 23, no. 10, pp. 1498–1512, 2011.

[22] N. Korfiatis, E. García-Bariocanal, and S. Sánchez-Alonso, "Evaluating content quality and helpfulness of online product reviews: The interplay of review helpfulness vs. review content," *Electronic Commerce Research and Applications*, vol. 11, no. 3, pp. 205–217, 2012.

[23] K. K. Kuan, K.-L. Hui, P. Prasarnphanich, and H.-Y. Lai, "What makes a review voted? an empirical investigation of review voting in online review systems," *Journal of the Association for Information Systems*, vol. 16, no. 1, p. 48, 2015.

[24] H. Baek, J. Ahn, and Y. Choi, "Helpfulness of online consumer reviews: Readers' objectives and review cues," *International Journal of Electronic Commerce*, vol. 17, no. 2, pp. 99–126, 2012.

[25] S. Lee and J. Y. Choeh, "Predicting the helpfulness of online reviews using multilayer perceptron neural networks," *Expert Systems with Applications*, vol. 41, no. 6, pp. 3041–3046, 2014.

[26] R. M. Schindler and B. Bickart, "Perceived helpfulness of online consumer reviews: the role of message content and style," *Journal of Consumer Behaviour*, vol. 11, no. 3, pp. 234–243, 2012.

[27] S. Lee and J. Y. Choeh, "The determinants of helpfulness of online reviews," *Behaviour & Information Technology*, pp. 1–11, 2016.

[28] M. Salehan and D. J. Kim, "Predicting the performance of online consumer reviews: A sentiment mining approach to big data analytics," *Decision Support Systems*, vol. 81, pp. 30–40, 2016.

[29] A. H. Huang, K. Chen, D. C. Yen, and T. P. Tran, "A study of factors that contribute to online review helpfulness," *Computers in Human Behavior*, vol. 48, pp. 17–27, 2015.

[30] S. M. Mudambi and D. Schuff, "What makes a helpful review? a study of customer reviews on amazon. com," *MIS quarterly*, vol. 34, no. 1, pp. 185–200, 2010.

[31] Y.-H. Cheng and H.-Y. Ho, "Social influence's impact on reader perceptions of online reviews," *Journal of Business Research*, vol. 68, no. 4, pp. 883–887, 2015.

[32] Y. Pan and J. Q. Zhang, "Born unequal: a study of the helpfulness of user-generated product reviews," *Journal of Retailing*, vol. 87, no. 4, pp. 598–612, 2011.

[33] S. Park and J. L. Nicolau, "Asymmetric effects of online consumer reviews," *Annals of Tourism Research*, vol. 50, pp. 67–83, 2015.

[34] L. M. Willemsen, P. C. Neijens, F. Bronner, and J. A. de Ridder, ""highly recommended!" the content characteristics and perceived usefulness of online consumer reviews," *Journal of Computer-Mediated Communication*, vol. 17, no. 1, pp. 19–38, 2011.

[35] S. Sen and D. Lerman, "Why are you telling me this? an examination into negative consumer reviews on the web," *Journal of interactive marketing*, vol. 21, no. 4, pp. 76–94, 2007.

[36] D. Yin, S. Bond, and H. Zhang, "Anxious or angry? effects of discrete emotions on the perceived helpfulness of online reviews," *Mis Quarterly*, vol. 38, no. 2, pp. 539–560, 2014.

[37] M. Li, L. Huang, C.-H. Tan, and K.-K. Wei, "Helpfulness of online product reviews as seen by consumers: Source and content features," *International Journal of Electronic Commerce*, vol. 17, no. 4, pp. 101–136, 2013.

[38] S. G. Moore, "Attitude predictability and helpfulness in online reviews: the role of explained actions and reactions," *Journal of Consumer Research*, vol. 42, no. 1, pp. 30–44, 2015.

[39] C. Forman, A. Ghose, and B. Wiesenfeld, "Examining the relationship between reviews and sales: The role of reviewer identity disclosure in electronic markets," *Information Systems Research*, vol. 19, no. 3, pp. 291–313, 2008.

[40] A. Qazi, K. B. S. Syed, R. G. Raj, E. Cambria, M. Tahir, and D. Alghazzawi, "A concept-level approach to the analysis of online review helpfulness," *Computers in Human Behavior*, vol. 58, pp. 75–81, 2016.

[41] K.-T. Lee and D.-M. Koo, "Evaluating right versus just evaluating online consumer reviews," *Computers in Human Behavior*, vol. 45, pp. 316–327, 2015.

[42] N. A. Ernst and G. C. Murphy, "Case studies in just-in-time requirements analysis," in *2012 Second IEEE International Workshop on Empirical Requirements Engineering (EmpiRE)*. IEEE, 2012, pp. 25–32.

[43] M. Glinz and S. A. Fricker, "On shared understanding in software engineering: an essay," *Computer Science-Research and Development*, vol. 30, no. 3-4, pp. 363–376, 2015.

[44] D. Yin, S. Mitra, and H. Zhang, "When do consumers value positive versus negative reviews? an empirical investigation of confirmation bias in online word of mouth," *Information Systems Research, forthcoming*, 2015.

[45] X. Li and L. M. Hitt, "Price effects in online product reviews: an analytical model and empirical analysis," *MIS quarterly*, pp. 809–831, 2010.

[46] N. Hu, L. Liu, and V. Sambamurthy, "Fraud detection in online consumer reviews," *Decision Support Systems*, vol. 50, no. 3, pp. 614–626, 2011.

[47] M. Luca and G. Zervas, "Fake it till you make it: Reputation, competition, and yelp review fraud," *Harvard Business School NOM Unit Working Paper*, no. 14-006, 2015.